Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/IL05/000263

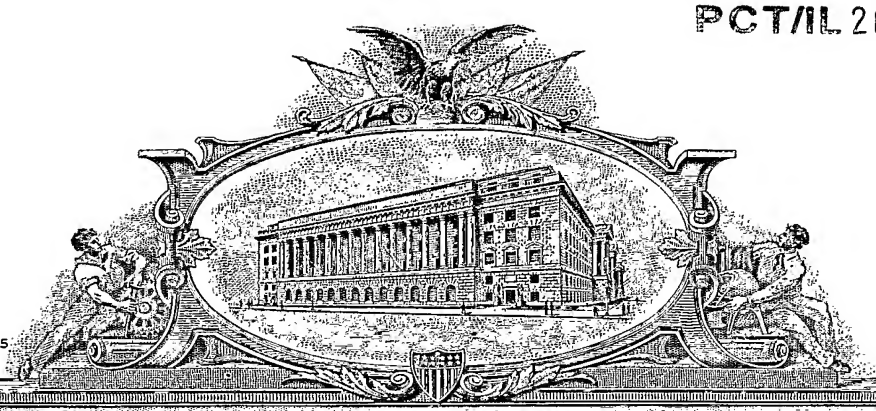
International filing date: 06 March 2005 (06.03.2005)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/549,541
Filing date: 04 March 2004 (04.03.2004)

Date of receipt at the International Bureau: 21 March 2005 (21.03.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



PA 1283735

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

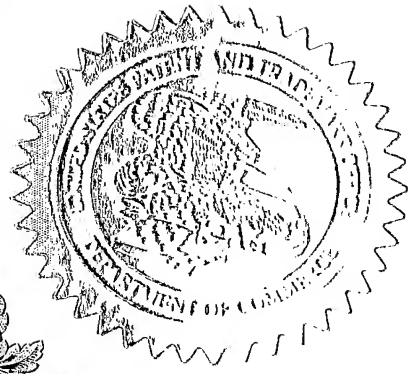
United States Patent and Trademark Office

February 15, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: ***60/549,541***
FILING DATE: ***March 04, 2004***

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS



W. MONTGOMERY
Certifying Officer

PTO/SB/16 (08-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No.

INVENTOR(S)		
Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
DENA	LESHKOWITZ	YEHUD, ISRAEL

Additional inventors are being named on the _____ separately numbered sheets *attached hereto*

TITLE OF THE INVENTION (500 characters max)

QUANTIFYING AND PROFILING ANTIBODY AND T CELL RECEPTOR GENE EXPRESSION

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☒ Customer Number: 24505

OR

☐ Firm or Individual Name	ALPHAPATENT ASSOCIATES LTD.				
Address	P.O.B. 2345				
Address					
City	BEIT SHEMESH	State		Zip	99544
Country	ISRAEL	Telephone	516-620-4572 (US)	Fax	516-620-4572 (US)

ENCLOSED APPLICATION PARTS (check all that apply)

☒ Specification *Number of* Pages 41 ☐ CD(s), Number _____

☒ Drawing(s) *Number of* Sheets 5 ☐ Other (specify) _____

☐ Application Date Sheet. See 37 CFR 1.76

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

☒ Applicant claims small entity status. See 37 CFR 1.27.

☐ A check or money order is enclosed to cover the filing fees.

☒ The Director is herby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 501380

☐ Payment by credit card. Form PTO-2038 is attached.

FILING FEE Amount ($): $80.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are: _____

Respectfully *submitted*,

SIGNATURE _____

TYPED or PRINTED NAME DANIEL J. SWIRSKY

TELEPHONE 516-620-4572 (US)

Date March 4, 2004

REGISTRATION NO. 45,148 (*if* appropriate)

Docket Number: 1300-USP

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

PTO/SB/17 (10-03)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

FEE TRANSMITTAL for FY 2004

Effective 1010112003. Patent fees are *subject to* annual revision.

☒ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 80.00

Complete if Known	
Application Number	
Filing Date	MARCH 4, 2004
First Named Inventor	Dena LESHKOWITZ
Examiner Name	
Art Unit	
Attorney Docket No.	1300-USP

METHOD OF PAYMENT (*check all* that apply)

☐ Check ☐ Credit card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:

Deposit Account Number: 501380

Deposit Account Name: ALPHAPATENT ASSOCIATES LTD.

The Director is authorized to: (*check all* that apply)

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, except for the filing fee to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code ($)	Small Entity Fee Code ($)	Fee Description	Fee Paid
1001 770	2001 385	Utility filing fee	80.00
1002 340	2002 170	Design filing fee	
1003 530	2003 265	Plant filing fee	
1004 770	2004 385	Reissue filing fee	
1005 160	2005 80	Provisional filing fee	

SUBTOTAL(1) ($) 80.00

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Fee from Extra Claims below	Fee Paid
Total Claims		-20- =		x
Independent Claims		- 3- =		x
Multiple Dependent				

Large Entity Fee Code ($)	Small Entity Fee Code ($)	Fee Description
1202 18	2202 9	Claims in excess of 20
1201 86	2201 43	Independent claims in excess of 3
1203 290	2203 145	Multiple dependent claim, if not paid
1204 86	2204 43	** Reissue independent claims over original patent
1205 18	2205 9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)

**or number previously paid, if *greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code ($)	Small Entity Fee Code ($)	Fee Description	Fee Paid
1051 130	2051 65	Surcharge -late filing fee or oath	
1052 50	2052 25	Surcharge - late provisional filing fee or cover sheet	
1053 130	1053 130	Non-English specification	
1812 2,520	1812 2,520	For filing a request for ex parte reexamination	
1804 920*	1804 920*	Requesting publication of SIR prior to Examiner action	
1805 1,840*	1805 1,840*	Requesting publication of SIR after Examiner action	
1251 110	2251 55	Extension for reply within first month	
1252 420	2252 210	Extension for reply within second month	
1253 950	2253 475	Extension for reply within third month	
1254 1,480	2254 740	Extension for reply within fourth month	
1255 2,010	2255 1,005	Extension for reply within fifth month	
1401 330	2401 165	Notice of Appeal	
1402 330	2402 165	Filing a brief in support of an appeal	
1403 290	2403 145	Request for oral hearing	
1451 1,510	1451 1,510	Petition to institute a public use proceeding	
1452 110	2452 55	Petition to revive - unavoidable	
1453 1,330	2453 665	Petition to revive - unintentional	
1501 1,330	2501 665	Utility issue fee (or reissue)	
1502 480	2502 240	Design issue fee	
1503 640	2503 320	Plant issue fee	
1460 130	1460 130	Petitions to the Commissioner	
1807 50	1807 50	Processing fee under 37 CFR 1.17(q)	
1806 180	1806 180	Submission of Information Disclosure Stmt	
8021 40	8021 40	Recording each patent assignment per property (times number of properties)	
1809 770	2809 385	Filing a submission after final rejection (37 CFR 1.129(a))	
1810 770	2810 385	For each additional invention to be examined (37 CFR 1.129(b))	
1801 770	2801 385	Request for Continued Examination (RCE)	
1802 900	1802 900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid SUBTOTAL (3) ($)

SUBMITTED BY

(Complete *(if applicable)*)

Name (Print/Type)	DANIEL J. SWIRSKY	Registration No. (Attorney/Agent)	45,148	Telephone	516-620-4572 (US)
Signature				Date	MARCH 4, 2004

QUANTIFYING AND PROFILING ANTIBODY AND T CELL RECEPTOR GENE EXPRESSION

INVENTOR: Dena Leshkowitz

FIELD OF THE INVENTION

The present invention relates to the sequencing of expressed genes belonging to the immunoglobulin gene superfamily, particularly immunoglobulins and T cell receptors. In particular, the present invention provides methods of sequencing expressed genes in a non-clonal population of cells of the immune system, generating profiles of the genes expressed, and correlating the data generated with states of disease or health.

BACKGROUND OF THE INVENTION

Immunoglobulins and T-cell Receptors

The immunoglobulin superfamily includes antibodies and T-cell receptors. Antibodies are antigen-binding proteins present on the B-cell membrane and secreted by effector B-cells called plasma cells. The serum antibodies produced in response to a particular antigen are heterogeneous due to the presence of multiple B-cell epitopes on antigens. These antibodies are termed polyclonal antibodies and their response facilitates the localization, phagocytosis, and the complement-mediated lysis of antigen.

An antibody structure consists of two identical heavy chains (H) and two identical light chains (L) which are linked by disulfide bonds. The amino terminal region in each heavy and light chain constitutes a variable sequence. Within the amino terminal variable domain of each heavy and light chain are three hypervariable regions, also called complementarity-determining regions (CDRs).

There are five antibody classes (IgM, IgG, IgD, IgA and IgE) and they differ in their ability to carry out various effector functions, in their average serum concentrations and their half-life.

In germ-line DNA, multiple gene segments encode a single immunoglobulin heavy or light chain. These genes segments are carried in the germ cells but can not be transcribed and translated into heavy and light chains until they are arranged into functional genes (Kuby J., Immunology, 3rd edition, 1997).

It has been shown in various studies that there is a restricted variable region usage in different developmental stages (Davidkova et al., Scand J Immunol, 45:62 (1997)), in malignant B cells (Kipps TJ et al., Proc Natl Acad Sci USA, 86:5913 (1989), in autoantibody repertoires (Pascual V et al., Int Rev Immunol, 5:231 (1990), in HCV infected individuals (Gasparotto D et al., Leukemia & Lymphoma, 43:747 (2002) and an "activated" repertoire different from the "normal" stable one (Dijk-Hard IV and Lundkvist I, Immunology 107:136, 2002). The common methods used for the above analysis is done by DNA or cDNA amplification by PCR and specific primers, cloning and sequencing. Due to the labor intensity involved, only a limited number of molecules were analyzed.

T-cell receptors (TCRs), unlike antibodies, do not react with soluble antigen but rather with processed antigen associated with a self-MHC molecule on an antigen-presenting cell or a target cell. The TCR variable region domains contain three hypervariable complementarity-determining regions, which appear to be equivalent to those in immunoglobulins. TCR germ-line DNA is organized into multigene families corresponding to the α, β, γ, and δ chains. Each multigene family contains multiple variable-region gene segments.

Current methods for selecting and producing monoclonal antibodies have been known in the art, and can provide the following aims: desired affinity, antigen-specificity and the isotype and with the desired effector function. Nonetheless monoclonal antibodies are produced in an artificial procedures including, but not limited, phage display technology (Clackson et al. Nature, 352:624-628,1991, Kretzschmar et al. Curr Opin Biotechnol. 13:598-602, 2002) and transgenic animals (Larry L. et al., U.S. Pat. No. 2003/0093820; Kellermann et al. Curr Opin Biotechnol., 13(6):593-597, 2002). These methods cannot imitate the exquisite versatility or complexity of the species genuine biological setting, lack the self-tolerance and repertoire shift, in particular the human immune system and in general non-human mammals.

DNA Microarray Technology

DNA microarrays, also known as "DNA Chips", are a potentially powerful technology for improving diagnostic classification, treatment selection and development of therapeutics.

In the past several years, this technology has emerged, promising to monitor the whole genome on a single chip, allowing better identification of drug protein candidates and gene expression profiling in different disease state conditions, especially in the diagnosis of a vast platform of malignancies for example in the diagnosis of haematological malignancies. It enables the analysis of RNA expression by clonal populations of leukemia and lymphoma cells on a genome-wide scale (Best Pract Res Clin Haematol. Dec; 16(4):645-52, 2003). In addition, monitoring multiple genes in parallel allows the identification of genes that are differentially expressed in malignant and normal tissues and classify theses genes as "signatures", of the disease state. Often, these signatures are impossible to obtain from tracking changes in the expression of individual genes, which can be subtle or variable.

Microarray for gene expression analysis is based on labeled cDNA or cRNA targets derived from the mRNA of an experimental sample are hybridized to nucleic acid probes attached to the solid support. By monitoring the amount of label associated with each DNA location, it is possible to infer the abundance of each mRNA species represented (de Saizieu, A., et al., Nature Biotech. 16: 45-48, 1998; and US Pat. No. 6,410,229). However, this technology requires prior knowledge of the gene sequence or customization for each new application. In the case of antibodies or TCRs due to the high variability of the sequence, this approach is not practical since every specimen will potentially be of interest.

Methods for Sequencing DNA

Although sequence information has already provided accelerated knowledge and potential resolution of diverse biological, medical and therapeutic research problems, in practice, actual sequencing requires large number of base pairs in order to obtain a reliable sequence.

Further challenges arise if sequencing projects are extended to include the determination of the genomic sequences of characteristic individuals or species of organisms, especially those that have economic, social or medical importance. Such sequencing projects would advance not only our understanding of the evolution of organisms and the evolution of biochemical processes, but would also further the detection, treatment and understanding of disease, and would aid agriculture, the food industry

and biotechnology in general. However beneficial the results of such projects would be, their successful completion requires the development of a new, rapid, reproducible and reliable sequencing method such as those described in this invention.

Accurate DNA sequencing is a crucial procedure in modern biological, medical and agricultural research. Traditionally DNA sequencing has been done by direct, base-by-base analysis, with each new base determination built on the results of many previous sequencing steps.

Direct sequencing techniques involve a variety of synthesis, degradation, or separation techniques, and include the traditional Sanger, pyrosequencing and exonuclease methods, as well as direct visualization approaches (see for example US patent application 2002/0009727; Shi, Clinical Chemistry 47:164-172, 2001; and Drmanac R., et al., Adv Biochem Engineering/Biotech 77:76-101, 2002).

In the Sanger method, DNA synthesis is randomly terminated at each base pair, creating a wide range of fragments that are then separated by gel according to length and scored. In the pyrosequencing method, polymerase-guided incorporation of each base is detected by measuring the pyrophosphate released in consecutive cycles. In the exonuclease approach, a single molecule of target DNA synthesized with fluorescent-tagged bases is degraded by exonuclease. The consecutively released nucleotides are then scored with a very sensitive fluorescence detector.

Indirect sequencing method such as "sequencing by hybridization" (SBH) is designed to sequence a target nucleic acid by allowing the identification of complementary sequences in the target nucleic acid by utilizing n-mers probes (all possible probes of length *n*). This method comprises the "universal microarray" approach, in which probes are designed using simple combinatorial and statistical principles, without guidance from prior knowledge of any specific gene or DNA sequence, was developed by several companies (Hyseq Inc., Genometrix Inc.) (Advances in Biochemical Engineering Biotechnology 77:76-101; and US Patent No. 5,695,940). The method relies on the process, in which oligonucleotide probes hybridize preferentially with entirely complementary and homologous nucleic acid targets are described. Using these hybridization conditions, overlapping oligonucleotide probes associate with a target nucleic acid. Following washes, positive hybridization signals are used to assemble the sequence of a given nucleic acid fragment. Representative target nucleic acids are applied as dots. Up to 4^8 or ~66,000 probes of the type (A,T,C,G)(A,T,C,G)N8(A,T,C,G) are used to determine sequence information by simultaneous hybridization with nucleic acid molecules bound to a filter. Additional hybridization

conditions are provided that allow stringent hybridization of 6-10 nucleotide long oligomers which extends the utility of the invention. A computer process determines the information sequence of the target nucleic acid which can include targets with the complexity of mammalian genomes. Sequence generation can be obtained for a large complex mammalian genome in a single process (U.S. Pat. No 5,525,464).

US Pat. No. 5,202,231, US Pat. No. 5,525,464 and WO 95/09248 disclose additional aspects of sequencing by hybridization. However, due to the repetitive nature of genomic DNA, this approach has not been proven as competitive as current biochemical sequencing methods. It has been used successfully for finding mutations (Gerry, et al J Mol Biol 292: 251-62, 1999). Various twists of this technology such as the use of universal bases are claiming improvement of the SBH process (Frieze, et al. J Comput Biol. 6: 361-368, 1999; Preparata FP et al., Comput Biol. 7: 621-30, 2000; and US Patent application 2003/0036073).

There is an unmet need for improved methods for sequencing and profiling expressed genes belonging to the immunoglobulin gene superfamily, particularly immunoglobulins and T cell receptors.

SUMMARY OF THE INVENTION

The present invention advantageously provides methods of sequencing expressed genes in a non-clonal population of cells of the immune system, generating profiles of the genes expressed, and correlating the data generated with states of disease or health.

The present invention provides methods that are suited to sequencing a plurality of target nucleic acids because they are capable of resolving sequence data prior to purifying the individual target nucleic acids. The present invention provides methods for sequencing a plurality of target nucleic acids in a mixture of unknown sequences by hybridization of overlapping short oligonucleotide probes of known or predicted sequence to the nucleic acid targets simultaneously. The oligomer probes of a given size can contain all or most existing combinations of nucleotides for complete sequencing and a part of all possible variants for partial sequencing. Probes can also be composed of oligomers of different sizes as well as comprising all known combinations of nucleotides that are possible for that size oligonucleotide.

According to one aspect of the invention, a plurality of oligonucleotides, each from eight to forty nucleotides in length is hybridized to the target sequences. Hybridization occurs using conditions which

are controlled and varied to ensure discrimination between perfectly matched oligonucleotides and oligonucleotides having a one base pair mismatch with the target sequence.

The present invention provides a method for sequencing a plurality of target nucleic acids in a mixture of unknown sequences comprising the steps of:

a) using conditions which differentiate an exactly complementary oligonucleotide probe and an oligonucleotide probe having at least one mismatched nucleotide;

(b) contacting a plurality of oligonucleotides, each from eight to forty nucleotides in length, with target nucleic acids;

(c) forming duplexes between the plurality of target nucleic acids and the plurality of oligonucleotides;

(d) washing the duplexes;

(e) detecting oligonucleotides positively hybridizing as part of said duplexes; and

(f) compiling a set of sequences of the plurality of nucleic acids from overlapping positively hybridizing oligonucleotides.

This invention provides methods for analyzing target nucleic acids from a single individual, selected from a group consisting of immunoglobulins and T cell receptor molecules, from a non-clonal or non-homogeneous source material. The invention further provides methods to generate the profile of sequences of expressed genes from such cell populations. The present invention further provides methods of generating databases from these data and correlating the expressed gene profile with states of disease or health.

Specifically, antibodies, surface or extracted immunoglobulins (Ig) and TCR are the targets in this invention. More specifically, this invention provides a method for identification, sequencing and quantification of the expression level or production level of target molecules and target profiling. Expression refers to the amount of a certain RNA molecule produced within a sample; production refers to the amount of specific target genes within the cell mixture.

According to one embodiment the present invention provides means for distinguishing the results obtained for different sequences in the target nucleic acids.

According to another embodiment the present invention provides an algorithm for analyzing the data obtained from the hybridization and eliminating hybridization errors.

According to yet another embodiment the present invention provides means for sorting the data in a data storage system.

The information procured by the methods of the present invention for the target molecules includes but is not limited to: class, subclass, family, partial or complete gene sequence and signature for a mixture of target molecules. Furthermore, the data obtained is stored in a data storage system.

The present invention further provides a method for partial sequencing of a plurality of nucleic acids comprising the steps of:

(a) using conditions which differentiate an exactly complementary oligonucleotide probe and an oligonucleotide probe having at least one mismatched nucleotide;

(b) contacting the plurality of target nucleic acids in a reaction mixture with a plurality, but less than a totality, of oligonucleotide probes of given length, each at least eight nucleotides in length;

c) forming duplexes between the target nucleic acids and the plurality of oligonucleotides;

(d) washing the duplexes;

(e) detecting oligonucleotides positively hybridizing as part of said duplexes; and

(f) compiling a partial sequence of said target nucleic acids from a subset of said oligonucleotide probes which form duplexes with said target nucleic acid and which overlap with at least one other member of said oligonucleotide probes.

In a preferred embodiment, the oligonucleotide probes can include every different probe that is complementary to a subsequence of the target nucleic acids. The probes can range from about 5 to about 80 nucleotides in length. More preferably the probes are from 8 to 40 in length. These probes are specifically designed in respect to germ line or other known target sequences. These longer and specific probes can be used in addition to the type of probes suggested in "universal microarray" approach

meaning short probes of all possible combinations. The invention limits itself to the target genes, however has the ability to discover and analyze any type of nucleic acid encoding for the target genes.

According to another aspect, the present invention provides a method of identifying nucleic acid probes to quantify the expression level of a plurality of target nucleic acids, comprising:

(a) providing a high density array of nucleic acid probes, comprising a multiplicity of nucleic acid probes, wherein each probe is complementary to a subsequence of a target nucleic acid and for each probe there is at least one corresponding mismatch control probe;

(b) hybridizing said target nucleic acid to said array of nucleic acid probes; and

(c) identifying those probes for which the difference in hybridization signal intensity between each probe and its mismatch controls is detectable.

The present invention is implemented utilizing known technologies including but not limited to "DNA Chips" as well as other types of microarrays. According to specific embodiments, the present invention uses probes that are conserved in the variable domain setting of the target genes. Each probe is complementary to a subsequence of a target nucleic acid and for each probe there is at least one corresponding mismatch control probe.

According to another aspect, the present invention provides a method of analyzing the expression of one or more genes encoding antibodies or antibody fragments and T cell receptors or T cell receptor fragments, comprising:

(a) providing a pool of target nucleic acids comprising RNA transcripts of one or more of said genes, or nucleic acids derived therefrom using said RNA transcripts as templates;

(b) hybridizing the pool of nucleic acids to an array of nucleic acid probes immobilized on a surface, said array comprising at least 100 different nucleic acids, at least some of which comprise control probes, wherein each different nucleic acid is localized in a known location on said surface, and at least some of said nucleic acid probes

are complementary to said RNA transcripts or said nucleic acids derived therefrom using said RNA transcripts; and

(c) quantifying the hybridization of said nucleic acids to said array by comparing binding of matched and control probes.

According to one embodiment, the control probes are mismatch control probes which are selected not to be perfectly complementary to a particular target nucleic acid sequence.

According to another embodiment, difference in hybridization signal intensity is calculated between each of said nucleic acid probes, and each corresponding mismatch control probes and its background signal.

According to another embodiment, the DNA array comprises preferably 10 different nucleic acid probes, more preferably 100 different nucleic acid probes, and most preferably 1000 different nucleic acid probes, complementary to subsequences of each gene.

According to one embodiment the target nucleic acids are extracted from cells of the immune system existing in any suitable specimen or sample, including but not limited to blood, tissue samples, selected cell subpopulations, bone marrow, lymph nodes, thoracic duct, Peyer's patch or any other organ or bodily fluid.

These and other embodiments of the present invention will become apparent in conjunction with the figures, description and claims that follow.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1- Flowchart of the invention process.

Figure 2- Compilation with overlapping positively labeled probes.

Figure 3 - Examples of multiple sample experiment.

Figure 4 - Scheme illustrating end labeled plus and minus strand.

Figure 5 – Scheme illustrating PCR amplification of specific regions within the target sequence.

DETAILED DESCRIPTION OF INVENTION

The term "biological sample", as used herein, refers to a sample obtained from an organism or from components (e.g., cells) of an organism. The sample may be of any biological tissue or fluid. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Such samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

The terms "nucleic acid" or "nucleic acid molecule" refer to a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, and unless otherwise limited, would encompass known analogs of natural nucleotides that can function in a similar manner as naturally occurring nucleotides.

An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 500 bases. Oligonucleotide probes are preferably from about 30 to about 100 nucleotides in length. Preferably, the oligonucleotide probes are from about 15 to 80 nucleotides in length, and more preferably from about 8 to about 40 nucleotides in length.

The term "target nucleic acid" refers to a nucleic acid derived from a biological sample to which the oligonucleotide probe is designed to specifically hybridize. It is either the presence or absence of the target nucleic acid that is to be detected, or the amount of the target nucleic acid that is to be quantified. The target nucleic acid has a sequence that is complementary to the nucleic acid sequence of the corresponding probe directed to the target. The term target nucleic acid may refer to the specific subsequence of a larger nucleic acid to which the probe is directed or to the overall sequence (e.g., gene or mRNA) whose expression level is to be detected.

The term "subsequence" refers to a sequence of nucleic acids that comprise a part of a longer sequence of nucleic acids.

As used herein a "probe" is defined as an oligonucleotide capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation .As used herein, an oligonucleotide probe may include natural (i.e., A, G, C, or T).

The term "stringent conditions" refers to conditions under which a probe will hybridize to its target subsequence, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. Typically, stringent conditions are those in which the salt concentration is at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30° C. Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding or non-specific interactions between the target nucleic acids and components of the oligonucleotide array (e.g., the oligonucleotide probes, control probes, the array substrate, etc.). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each target nucleic acid.

The term "mismatch control" refers to a probe that has a sequence deliberately selected not to be perfectly complementary to a particular target sequence. The mismatch control typically has a corresponding test probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases. While the mismatch(s) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely destabilizes the duplex with the target sequence under the test hybridization conditions.

The term "quantifying" when used in the context of quantifying transcription levels of a gene can refer to absolute or to relative quantification.

The present invention relates to gene sequence analyses using various target gene regions. Specifically, the variable regions such as the frameworks, complementarity-determining regions (CDRs), diversity (D), joining (J), constant (C) and leader regions or any other region within the target genes. The analysis of these regions enables partial or full target genes reconstruction, classification, quantification and profiling.

The immune system has several salient features which are relevant for this present invention:

Diversity

The number of different antigen receptors expressed by mature lymphocytes is estimated in excess of 10^9 for T and B cells. This diversity enables the immune system to defend against a broad range of antigens. The diversity is a result of multiple germ line genes which undergo recombination and mutations.

The antibodies and T cell receptors (TCRs) contain segments that vary greatly among the molecules types, these segments are termed hypervariable region (V region). The specificity of an antibody or TCR for an antigen is determined by three complementary determining regions (CDR) in the V region of both heavy (H) and light (L) chains. The "variable region" defines the binding specificity, more specifically the antigen-binding site. The sequence of the variable region is approximately 330 bases in length. The variable sequence is composed of regions that are conserved among the immunoglobulin chains denoted framework regions. The regions that are highly variable, i.e., the CDRs appear at predetermined intervals within the relatively constant framework residues.

Combinatorial association of immunoglobulin gene elements is the most important process in the creation of extreme diversity of antibody molecules. The recombination of germ-line variable gene elements V, D, and J (51 gene elements for Vh, 27 gene elements for D, 6 gene elements for Jh, 40 gene elements for Vkappa, 5 gene elements for Jkappa, 30 gene elements for Vlambda, 4 gene elements for Jlambda) can potentially generate approximately 6000 variable genes of human heavy chains. As joining of these elements is imprecise and is occurring, for example, with nucleotide insertions or deletions, the created diversity is in fact much higher. The assembled variable genes can be revised and edited resulting in a change of their affinity and even specificity.

Self-tolerance

Self-tolerance is a phenomenon whereby clones expressing any potential anti-self antibodies are eliminated. Thus, under normal circumstances the repertoire of cells will not include clones expressing an Ig or TCR complementary to the organism normal self-molecules.

Repertoire Shift

During T cell dependent immune responses, the average affinity for the eliciting antigen increases with time. This affinity maturation is generated by somatic hypermutation and selection for B cells

carrying high-affinity immunoglobulins. A secondary immune response is characterized by rapid differentiation of memory B cells into plasma cells producing high affinity antibodies. However, the V gene repertoire of a secondary response differs from that used in the primary response. This phenomenon is called repertoire shift.

The present invention belongs to the field of molecular biology. It involves a novel method of sequencing of a target nucleic acid sequence by hybridization of short oligonucleotide probes to a nucleic acid target. The oligonucleotide probes can comprise all known combinations of the four nucleotides of a given length, i.e. oligonucleotides of base composition adenine (A), thymine (T), guanine (G), and cytosine (C) for DNA and A,G,C, and uridine (U) for RNA. Conditions are described which allow hybridization discrimination between oligonucleotides which are as short as six nucleotides long and have a single base end-mismatch with the target sequence .

Using RNA as The Biological Material Source

There is a certain advantage in using RNA for the analyses of the invention. If genomic DNA is used, one needs to differ between the two alleles for the target genes, since only one is expressed and rearranged. In addition the invention is aimed to cells expressing the target cells and not to non-mature forms of the targets. Therefore, when using DNA as the biological material, previous enrichment for cells expressing the target genes or CD markers should be considered. This enrichment will select the cells that contain rearranged target gene.

Furthermore, PCR specific primers are selected that will amplify the specific J region that is most adjacent to the V region as opposed to downstream J segments. However, if RNA is used, the RT with the oligo dT primer will select only polyA RNA for cDNA creation and therefore only the active J region will be present in the amplified sequence.

DNA Microarrays

Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (e.g. through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. Thus, changes in the expression (transcription) levels of particular genes, especially genes that encode immunoglobulins and TCRs, serve as signposts for the presence and progression of various

diseases including but not limited to immunodeficiencies, cancers (e.g. lymphomas and leukemia), viral infection and autoimmune diseases.

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid) and have been used to detect expression of particular genes (e.g., a Northern Blot). In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose (See U.S. Pat. Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126).

Arrays of probes spotted onto the surface of modified microscope glass slides are described in Shena et al., Science 270: 467-470, 1995; and Shalon et al., Genome Research 6: 639-645, 1996. Arrays in which the probes have been grown on the surface of a substrate are described in Lockhart et al., Nature Biotechnology 14:1675, 1996.

U.S. Pat. Nos. 6,040,138 and 6,410,229 disclose methods of monitoring the expression levels of a multiplicity of genes. The methods involve hybridizing a nucleic acid sample to a high density array of oligonucleotide probes where the high density array contains oligonucleotide probes complementary to subsequences of labeled target nucleic acids in the nucleic acid sample and quantifying the hybridization of nucleic acids to the array. U.S. Pat. No. 6,548,257 provides methods for identifying nucleic acid probes to quantify the expression of a target nucleic acid.

Means of detecting such labels are well known to those of skill in the art. The label may be added to the target (sample) nucleic acid(s) prior to, or after the hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an avidin-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected. The indirect method using amino allyl dUTP is

preferred compared to the direct method using fluorescent probes. For a detailed review of methods of labeling nucleic acids and detecting labeled hybridized nucleic acids see Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., 1993.

The present invention provides a DNA microarray that contains oligos covering the junctions between the gene fragments composing the variable region. Tables 1 and 3 demonstrate the various combinatorial possibilities for the rearranged variable region. In the case of the light Ig chain, the rearrangement is of the joining of VJ gene segments (see Table 1). In the case of the heavy chain the rearrangement is of the joining of VDJ gene segments (see Table 3). Variable number of bases can be deleted or added (0-15) at the junction. There are several options in revealing the junctions sequence (one junction in the light chain and two junctions in the heavy chain):

1. Using a "true" universal chip that contains all possible combinations of n-mers.
2. Selecting the more abundant representative sequences and for them designing specific PCR primers, the PCR product can be sequenced by conventional methods.
3. The junction can be revealed by oligos that posses partial overlap with the germline gene flanked by random sequence. This is illustrated in Table 6, where N symbolizes any one of the bases (ATGC).

In addition, the DNA chip may advantageously contain positive and negative controls such as various primers used in the RT and PCRs and Spike controls. Spike controls represent any DNA that will not bind to the target designed probes for instance a poly run of a single nucleotide. Positive control can be a spot containing all possible combination probes.

Focusing on the Variable Region or CDR Regions

An important aspect of the invention is the focus on the variable region and CDR regions as experimental targets. The labeled target for the hybridization in this invention encodes for antibodies and TCR and mainly focuses on the variable region. This specification is done by PCR or RT-PCR reactions using specific primers.

Suitable Oligonucleotide Probes and Targets

A mismatch-free hybridization of oligonucleotides to unknown target nucleic acids represents, in essence, a sequencing of complementary target. Probability calculations and, in part, computer simulations estimate the types and numbers of oligonucleotides that would have to be synthesized in order to sequence a megabase plus segment of DNA .

In order to obtain the required amount of experimental data, the number of probes can be reduced depending on the number of fragments used and vice versa. The useful probes are those longer than 6 bases, since the shorter ones would require use of unsuitable and unpractical nucleic acid fragments shorter than a few hundred bases long.

There are two parameters which influence the choice of probe length. The first is the success in obtaining hybridization results that show the required degree of discrimination. The second is the technological feasibility of synthesis of the required number of probes .

The requirement of obtaining sufficient hybridization discrimination with practical and useful amounts of target nucleic acid limits the probe length from both sides. It is difficult to obtain a sufficient amount of hybrid with short probes, and to discriminate end mismatch with long probes. There is no evidence for use of probes shorter than 11-mers in the literature, except for very stable ones (Estivill et al., Nucl. Acids Res. 15: 1415-1425, 1987). On the other hand, probes longer than 15 bases discriminate end mismatch with difficulty (Wood et al., Proc. Natl. Acad. Sci. USA 82: 1585-1588, 1985) .

One solution for the problem of unstable probes and end mismatch discrimination is the use of a group of longer probes representing a single shorter probe in an informational sense. Groups of sixty-four 11-mers can be used instead of single 8-mers. Every member of the group has a common core 8-mer and one of three possible variations on outer positions with two variations at each end. The probe can be represented as 5'(A,T,C,G)(A,T,C,G)N8(A,T,C,G) 3'. With this type of probe one does not need to discriminate the non-informative end bases (two on 5' end, and one on 3' end) since only the internal 8-mer is read. This solution requires the use of higher mass amounts of probes in hybridization reactions.

The present invention provides designed chip which contains a set of oligos that is derived from the germline gene segments and in addition contains oligos that have a single mutation at each one of the oligo bases (see Table 1). As the length of the oligo used is longer there is a greater need to insert more

mutations. Since the variance of a certain antibody from the germline sequence is about 10%, when using 8 mers it is sufficient to use single mutations. However, if 12 mers are used than the oligo set should contain oligos with both a single mutation and oligos with two mutations. It should be noted that even if a certain 12 base sequence within the variable has more than 2 mutations, then the correct sequence might still be revealed, assuming that the three mutations are not juxtaposed (see Example 12E). The amount of mutation in the region of CDRs is higher than in the region of the frameworks and in addition the CDRs compose only ~1/3 of the variable region. Therefore, the chip contains more variants of mutation for the region of the CDRs, such as 12 mers oligos that in the framework will be designed to contain 0-2 mutations whereas for the CDRs the oligos will be designed to contain 0-4 mutations.

According to the present invention the final construction of the variable region of the target genes in a sample is achieved by a two-step protocol. Basically, in the first step only the frameworks family or the framework complete sequences are determined. In the second step, the CDR regions are revealed. The first step uses an array containing longer oligos (10-30 bases) designed to match framework sequence region. The selected oligos are designed by similar Tm. They will contain the germline sequence and mismatches (one or two per oligonucleotide). The hybridization results from this framework array will be used in the second step. In this step the most predominant and the more "certain" frameworks will be used to design specific primers that will be used in the PCR (as shown in Figure 4 and Figure 5). This step focuses us at the more significant data and reduces the amount of signals for the CDR sequence reconstruction. The amount of granulation needed, meaning the number of PCRs and arrays to be used in the second step depends on the complexity of the sample as revealed by the first step .The ultimate goal is to reveal the protein sequence, therefore the redundant DNA sequences that encode the same protein sequence can be merged to a single probe with degenerate nucleotides or universal nucleotides according to the amino acid codon table (see Example 13). Using this approach the probes should be longer than in a probes set were wobble degeneracy is not implemented in order to reduce the amount of probes that appear several times in a sequence or that are shared between different sequences.

Hybridization

Hybridization depends on the pairing of complementary bases in nucleic acids and is a specific tool useful for the general recognition of informational polymers. Diverse research problems using hybridization of synthetic oligonucleotide probes of known sequence include, amongst others, the

different techniques of identification of specific clones from cDNA and genomic libraries; detecting single base pair polymorphisms in DNA; generation of mutations by oligonucleotide mutagenesis; and the amplification of nucleic acids in vitro from a single sperm, an extinct organism, or a single virus infecting a single cell.

It is possible to discriminate perfect hybrids from those hybrids containing a single internal mismatch using oligonucleotides 11 to 20 nucleotides in length (Wallace et al., Nucl. Acids Res. 6: 3543-3557, 1979). Mismatched hybrids are distinguished on the basis of the difference in the amount of hybrid formed in the hybridization step and/or the amount remaining after the washing steps (Ikuta et al., Nucl. Acids Res. 15: 797-811, 1987; Thein and Wallace, in Human Genetic Diseases: A Practical Approach, ed. by J. Davies, IRL Press Ltd., Oxford, 33-50, 1986).

The reproducible hybridization of different and diverse short oligonucleotides less than 11 nucleotides long has not been well characterized previously. Detailed hybridization data that allows a constant set of conditions for all predictable oligonucleotides is not available (Besmer et al., J. Mol. Biol. 72: 503-522, 1972; Smith, Methods of DNA and RNA Sequencing, ed. S. Weissman, Praeger Publishers, New York, N.Y., 23-68, 1983; Estivill et al., Nucl. Acids Res. 15: 1415-1425, 1987).

Information is also not available on the effects of a single noncomplementary base pair located at the 5' or 3' end of a hybridizing oligonucleotide that produces a mismatched hybrid when associated with a target nucleic acid. Hybridization conditions that discriminate between (1) a perfectly complementary hybridizing pair of nucleic acid sequences where one partner of the pair is a short oligonucleotide, and (2) a pair wherein a mismatch of one nucleotide occurs on the 5' or 3' end of the oligonucleotide, provide a more stringent environment than is required for internal mismatches because hybrid stability is affected less by a mismatch at the end of a hybridizing pair of complementary nucleic acids than for an internal mismatch.

The length of nucleotides that can distinguish a unique sequence in a nucleic acid of defined size has been predicted (Smith in Methods of DNA and RNA Sequencing, ed. S. Weissman, Praeger Publishers, New York, N.Y., 23-68, 1983). Thus random oligonucleotide sequences 16-17 long are expected to occur only once in random DNA $3X10^9$, the size of the human genome. However, with decreasing probe length, e.g. for oligonucleotides 5 to 10 nucleotides in length, there is an exponential increase in the frequency of occurrence within a random DNA of a given size and complexity. Thus, the

purposes for which oligonucleotide probes are employed can impact on the length of the oligonucleotides that are used experimentally.

Plus/Minus Strand Approach

According to the present invention, the two labeled DNA strands are hybridized to the array.

Depending on the experimental design, the hybridization can be to the same array or to two duplicated arrays. If the hybridization is to the same array, two different labels are used such as cy3 and cy5 – one for labeling the plus strand and the other for the minus strand (Figure 4). This approach has several advantages:

a) Each sequence will be confirmed since both plus and minus strands are used.

b) There is a reduction of the impact of false negative and of the impact of false positive. False negative is the case that oligos that should have produced a significant hybridization signal and did not. False positive is the case that oligos that produced a significant hybridization signal yet are not complementary to target sequence. The likelihood of false positive and false negative is, overall, reduced since they need to insert the same effect on the plus and minus strands, in order to cause a real problem.

c) This experimental setting, allows oligos to function as internal controls to themselves. Thus, probes that bind weakly to the target nucleic acids will also have a weak non-specific binding with the opposite strand.

Inevitable uncertainties in the hybridization process caused for example, by the appearance of a probe in the both plus and minus strand will be disregarded since one skilled in the art analyzes the difference between the hybridization signal of the plus and minus strands. The solution is to give a "present" definition for oligos that have a significantly high value of hybridization for both strands and absent.

Conditions For Hybridization Stringency

Wallace et al. (Nucl. Acids Res. 6: 3543-3557,1979) describes conditions that differentiate the hybridization of 11 to 17 base long oligonucleotide probes that match perfectly and are completely homologous to the target nucleic acid as compared to similar oligonucleotide probes that contain a single internal base pair mismatch. Wood et al. (Proc. Natl. Acad. Sci. 82: 1585-1588, 1985) describes conditions for hybridization of 11 to 20 base long oligonucleotides using 3M-tetramethyl ammonium chloride wherein the melting point of the hybrid depends only on the length of the oligonucleotide probe, regardless of its GC content. However, as disclosed in these references eleven mer oligonucleotides are the shortest ones that generally can be hybridized successfully, reliably and reproducibly using known hybridization conditions.

Hybridization Reference

The hybridization pattern of the various individual mutligene germline (VDJ) is studied and used as reference. The hybridization pattern from the multigenes will give an estimate on probes that are false positive and false negative. Given a hybridization pattern of a sample, the variation of this pattern from the multigene pattern will be studied in order to deduce the sequences that are specific to the sample. The hybridization data achieved is the basis for computational models such as training neural networking processes. This approach has been demonstrated in predicting the sequence of TP53 gene based on p53 GeneChip (Spicker JS et al. Bioinformatics 18, 1133-1134, 2002).

Sequencing

Nucleic acid sequencing methods, where the position of each base in a nucleic acid molecule in relation to its neighbors is determined to define its primary structure, were developed in the early 1960's for RNA molecules and in the late 1970's for DNA. The two major methods for DNA sequencing, i.e. chemical degradation and dideoxy-chain termination, involve identification and characterization of 1-500 nucleotide long DNA fragments, specific for each one of at least four nucleotide bases, on polyacrylamide gels. The polyacrylamide gels must be able to distinguish single base pair differences between fragments. The fragments are generated either by chemical degradation (Maxam-Gilbert, Proc. Natl. Acad. Sci. 74: 560-564, 1977) or by dideoxy-chain termination of DNA fragments synthesized by DNA polymerase (Sanger et al., Proc. Natl. Acad. Sci. 74: 5463-5467, 1977). A sufficient quantity of isolated fragments is

ensured by recombinant DNA technology methods which include cloning, restriction enzyme digestion, gel electrophoresis, and polymerase chain reaction amongst others. These methods allow the identification and amplification of the target DNA to provide material for sequencing.

Although the use of electronic robots and computers allow acceleration of the number of base pairs actually determined, preparation of samples, electrophoresis and the subsequent manipulations necessary to obtain high quality reads by machines still involve significant intensive, skilled, manual labor for which no substitutes have been found.

Determining Sequence from The Oligonucleotide Hybridization Data

According to the present invention, of a plurality of target nucleic acids can be visualized as consisting of two steps;

1) Initially there is an evaluation process for the existence (production) of the germline genes taking into consideration all the combinatorial possibilities of the germline gene fragments rearrangement. By summing up the value of all the probes that indicates the presence of these genes one skilled in the art can evaluate whether a certain transcript has a positive signal .

2) All the predominant combinatorial events, in other words for the high scoring combinations go through a fine tuning step in order to score the various mutations and junctions.

In the invention, hybridization of all possible N-mer oligonucleotide probes to the target nucleic acids determines the N-mer oligonucleotides subsets contained in the primary sequence of the target nucleic acids and is the first step in the process of the invention. Positively hybridizing N-mer oligonucleotide probes are ordered and the sequence of the target DNA is determined using (N-1)mer (or less) overlapping frames between the oligonucleotide probes.

Reassembling the sequence of known oligonucleotides that hybridize to the target nucleic acids to generate the sequence of the target nucleic acid cannot be accomplished in some cases. This is because some information is lost if the target nucleic acid is not in fragments of appropriate size in relation to the size of oligonucleotide that is used for hybridizing. The quantity of information lost is proportional to the

length of a target being sequenced. However, if sufficiently short targets are used, their sequence can be unambiguously determined.

The probable frequency of duplicated sequences that would interfere with sequence assembly which is distributed along a certain length of DNA can be calculated. This derivation requires the introduction of the definition of a parameter having to do with sequence organization: the sequence subfragment (SF). A sequence subfragment results if any part of the sequence of a target nucleic acid starts and ends with a (N-1)mer that is repeated two or more times within the target sequence. Thus, subfragments are sequences generated between two points of branching in the process of assembly of the sequences in the method of the invention.

In this invention the branching problem is less pronounced due to the special sequence structural patterns for alternating frameworks and CDRs in the target molecules. This structure suggests an advantage since the highly variable regions are fairly short and the (many) junctions with the conserved regions embed information that facilitates the sequence assembly. The conserved regions serve as anchors since their location in the gene and their and composition confines the possibilities need to be explored in the sequence discovery process.

Data base Creation and Information Obtained Using the Method

Creating a database containing the clinical patient information along with the results of the biological sample including the: target gene profile, expression and sequence information. Crossing this information among multiple samples and populations can reveal important novel antibodies and TCRs along with correlation to development, health condition.

1. Oligos Profile:

   This output reveals the individual oligos hybridization intensity as revealed after preprocessing steps. This output can be used in analysis involving multiple samples in order to find similarities and/or differences.

2. Germline Profile:

   This output reveals the separate germline gene fragments expression level. The junctions are not evident in this output.

3. Complete Chain Profile:

This output reveals the complete variable regions existing within the sample with their relative amounts. This output contains both the nucleotide and the protein sequence information. The protein is determined by the translation of the nucleotide sequences. In addition it can identify nucleotide sequences that can produce a fully active antibody.

The present invention provides unique advantages compared to other methods known in the art:

1. High Throughput: this method allows analysis of a large amount of cells and samples in a fast manner in order to identify a number of relevant target nucleic acids. Analyze a mixture of a plurality of nucleic acids: this method allows simultaneous analysis of a plurality of nucleic acids, meaning it is not restricted to analysis of a single target molecule.

2. Multiple population analysis: ability to easily screen a broad spectrum of samples, considering the disease stage, developmental stage, type of treatment and other measures. This supports the ability to identify a number of relevant targets of greatest effectiveness prospects.

3. Analyze the quantitative level of expression of the various target nucleic acids: this ability is achieved by analyzing the extent of hybridization to multiple probes defining the target nucleic acids. There is evidence showing that the V family usage is restricted and therefore a mixture of targets found within a sample can have a commonality or in other words they can have a restricted repertoire. Furthermore, this repertoire can be significantly different from a second sample

4. Integrated discovery of novel targets: ability to identify their sequence or partial sequence and/or classify the target gene as well as their expression profile. This allows concentrating on targets with a significant profile, meaning higher correlation with population sub-type and ability to scan for the more pronounced targets .

5. Ability to discover human biological targets. The advantages of the "true" targets are that they have undergone selective pressure that: restricts them from identifying self and insures high affinity and specific binding to antigen.

6. Shortening of target discovery and antibody manufacture time: Currently examining the repertoire of target genes in a diverse sample requires cloning of the PCR products and then individual clone picking and sequencing which is labor and time demanding (a few days). The extent of clone picking and sequencing is proportional to the amount of complexity within a sample. However, this information is not known a priori. In addition in order to infer on the expression level of a certain sequence within the sample one needs to perform massive analysis. These steps are shortened to a days work using this application method.

7. Ability to simultaneously explore both TCR and Ig genes: Information on both these target types within the same sample can result in a more comprehensive picture of the immune response .

8. Obtaining a signature or profile of the target genes within populations: The amount of hybridization between the targets in the sample and the probe set taken as a whole composes a profile or signature. The study of populations profile can reveal their environmental, disease or any other influencing factor history. This information could be used for diagnostic and research purposes.

The following examples are to be considered merely as illustrative and non-limiting in nature. It will be apparent to one skilled in the art to which the present invention pertains that many modifications, permutations, and variations may be made without departing from the scope of the invention.

EXAMPLES

Example 1- Table 1: Representation of The Ig Lambda Chain Gene Segments

#	Sub Family	Locus	Sub Family
1	VL1	1a	JL1
2	VL1	1e	JL2
3	VL1	1c	JL3
4	VL1	1g	JL4
5	VL2	2c	JL5
6	VL2	2e	JL6
7	VL2	2a2	JL7
8	VL2	2d	
9	VL3	3r	
10	VL3	3j	
11	VL3	3p	
12	VL3	3a	
13	VL3	3l	
14	VL3	3e	
15	VL3	3m	
16	VL3	2-19	
17	VL4	4c	
18	VL4	4b	
19	VL5	5e	
20	VL5	5c	
21	VL5	5b	
22	VL7	7a	
23	VL7	7b	
24	VL8	8a	
25	VL9	9a	
26	VL10	10a	

5

Example 2

Table 2 - All Possible Oligos Derived from VH1-18 (Base 1 to 12) Containing Zero-One Mutations

Gene fragment	Start position in gene	# Mutations (position in oligo,type)	Oligo sequence
VH1-18	1	0	CAGGTTCAGCTG
VH1-18	1		
VH1-18	1	1 (1, c->g)	gAGGTTCAGCTG
VH1-18	1	1 (1, c->a)	aAGGTTCAGCTG
VH1-18	1	1 (1, c->t)	tAGGTTCAGCTG
VH1-18	1	1 (2, a->c)	CcGGTTCAGCTG
VH1-18	1	1 (2, a->g)	CgGGTTCAGCTG
VH1-18	1	1 (2, a->t)	CtGGTTCAGCTG
VH1-18	1	1 (3, g->c)	CAcGTTCAGCTG
VH1-18	1	1 (3, g->a)	CAaGTTCAGCTG
VH1-18	1	1 (3, g->t)	CAtGTTCAGCTG
VH1-18	1	1 (4, g->c)	CAGcTTCAGCTG
VH1-18	1	1 (4, g->a)	CAGaTTCAGCTG
VH1-18	1	1 (4, g->t)	CAGtTTCAGCTG
VH1-18	1	1 (5, t->a)	CAGGaTCAGCTG
VH1-18	1	1 (5, t->c)	CAGGcTCAGCTG
VH1-18	1	1 (5, t->g)	CAGGgTCAGCTG
VH1-18	1	1 (6, t->a)	CAGGTaCAGCTG
VH1-18	1	1 (6, t->c)	CAGGTcCAGCTG
VH1-18	1	1 (6, t->g)	CAGGTgCAGCTG
VH1-18	1		
VH1-18	1	1 (7, c->a)	CAGGTTaAGCTG
VH1-18	1	1 (7, c->t)	CAGGTTtAGCTG
VH1-18	1	1 (7, c->g)	CAGGTTgAGCTG
VH1-18	1	1 (8, a->t)	CAGGTTCtGCTG
VH1-18	1	1 (8, a->c)	CAGGTTCcGCTG
VH1-18	1	1 (8, a->g)	CAGGTTCgGCTG
VH1-18	1	1 (9, g->c)	CAGGTTCAcCTG
VH1-18	1	1 (9, g->a)	CAGGTTCAaCTG
VH1-18	1	1 (9, g->t)	CAGGTTCAtCTG
VH1-18	1	1 (10, c->g)	CAGGTTCAGgTG
VH1-18	1	1 (10, c->a)	CAGGTTCAGaTG
VH1-18	1	1 (10, c->t)	CAGGTTCAGtTG
VH1-18	1	1 (11, t->a)	CAGGTTCAGCaG
VH1-18	1	1 (11, t->c)	CAGGTTCAGCcG
VH1-18	1	1 (11, t->g)	CAGGTTCAGCgG
VH1-18	1	1 (12, g->a)	CAGGTTCAGCTa
VH1-18	1	1 (12, g->c)	CAGGTTCAGCTc
VH1-18	1	1 (12, g->t)	CAGGTTCAGCTt

Example 3

Table 3 - Representation of the Ig Heavy Chain Gene Segments

Variable germline gene fragment recombined with D and then J gene segments. N (0-15) random bases are inserted in the junctions between the segments. The arrows represent a certain recombination event joining VH1, 1-03 with D2, 2-21 and JH4 (V- (N) - D - (N) – J).

#	Sub Family	Locus		Sub Family	Locus		Sub Family
1	VH1	1-02		D1	1-1		JH1
2	VH1	1-03		D1	1-7		JH2
3	VH1	1-08		D1	1-20		JH3
4	VH1	1-18		D1	1-26		JH4
5	VH1	1-24		D2	2-2		JH5
6	VH1	1-45		D2	2-8		JH6
7	VH1	1-46		D2	2-15		
8	VH1	1-58		D2	2-21		
9	VH1	1-69		D3	3-3		
10	VH1	1-e		D3	3-9		
11	VH1	1-f		D3	3-10		
12	VH2	2-05		D3	3-16		
13	VH2	2-26		D3	3-22		
14	VH2	2-70		D4	4-4		
15	VH3	3-07		D4	4-11		
16	VH3	3-09		D4	4-17		
17	VH3	3-11		D4	4-23		
18	VH3	3-13		D5	5-5		
19	VH3	3-15		D5	5-12		
20	VH3	3-20		D5	5-18		
21	VH3	3-21		D5	5-24		
22	VH3	3-23		D6	6-6		
23	VH3	3-30		D6	6-13		
24	VH3	3-30.3		D7	7-27		
25	VH3	3-30.5					
26	VH3	3-33					
27	VH3	3-43					
28	VH3	3-48					
29	VH3	3-49					
30	VH3	3-53					
31	VH3	3-64					
32	VH3	3-66					
33	VH3	3-72					
34	VH3	3-73					
35	VH3	3-74					
36	VH3	3-d					

37	VH4	4-04					
38	VH4	4-28					
39	VH4	4-30.1					
40	VH4	4-30.2					
41	VH4	4-30.4					
42	VH4	4-31					
43	VH4	4-34					
44	VH4	4-39					
45	VH4	4-59					
46	VH4	4-61					
47	VH4	4-b					
48	VH5	5-51					
49	VH5	5-a					
50	VH6	6-01					
51	VH7	7-4.1					

Example 4

Table 4 - Scoring the Various Sequence Reconstruction Options

		Sequence	Score Example 1	Score Example 2
a	D2-15 .	AGGATATTGTAGTGGTGGTAGCTGCTACTCC	80	80
b	-1------2-	AGGgTATTGTAGTGGTGGTA[redacted]TGCTACTCC	100	80
c	-1--------	AGGgTATTGTAGTGGTGGTA[redacted]TGCTACTCC	85	90
d	--------2-	AGGATATTGTAGTGGTGGTA[redacted]TGCTACTCC	90	90

The example in Table 4 illustrates how to decide which of the sequences coexist in the sample. If the double mutation scores higher than any of the individual ones then it is mostly likely that it is the only predominate one, However, if each of the single mutations scores higher than the double one than it is most likely that both single mutations exist and not the double mutation. In this particular case a more definite result could be obtained by using oligos with 20 bases.

Example 5

Table 5 - An Example for the Percent Expression of the Germline Gene Fragments Within a Sample

#	Sub Family	Locus	% Expression	Sub Family	Locus	% Expression	Sub Family	% Expression
1	VH1	1-02		D1	1-1		JH1	
2	VH1	1-03		D1	1-7	50	JH2	40
3	VH1	1-08		D1	1-20		JH3	
4	VH1	1-18		D1	1-26		JH4	60
5	VH1	1-24		D2	2-2	10	JH5	
6	VH1	1-45	40	D2	2-8			
7	VH1	1-46		D2	2-15			
8	VH1	1-58		D2	2-21	40		
9	VH1	1-69		D3	3-3			
10	VH1	1-c		D3	3-9			
11	VH1	1-f		D3	3-10			
12	VH2	2-05		D3	3-16			
13	VH2	2-26		D3	3-22			
14	VH2	2-70		D4	4-4			
15	VH3	3-07	10	D4	4-11			
16	VH3	3-09		D4	4-17			
17	VH3	3-11		D4	4-23			
18	VH3	3-13		D5	5-5			
19	VH3	3-15		D5	5-12			
20	VH3	3-20		D5	5-18			
21	VH3	3-21		D5	5-24			
22	VH3	3-23		D6	6-6			
23	VH3	3-30		D6	6-13			
24	VH3	3-30.3						
25	VH3	3-30.5						
26	VH3	3-33						
27	VH3	3-43						
28	VH3	3-48						
29	VH3	3-49						
30	VH3	3-53						
31	VH3	3-64						
32	VH3	3-66						
33	VH3	3-72						
34	VH3	3-73						
35	VH3	3-74						
36	VH3	3-d						
37	VH4	4-04						
38	VH4	4-28						
39	VH4	4-30.1						
40	VH4	4-30.2						
41	VH4	4-30.4						
42	VH4	4-31						
43	VH4	4-34	50					
44	VH4	4-39						
45	VH4	4-59						
46	VH4	4-61						
47	VH4	4-b						

48	VH5	5-51						
49	VH5	5-a						
50	VH6	6-01						
51	VH7	7-4.1						

Example 6 - Oligos Designed for the V-D Junction

An example of the oligos is selected to span the VH7-4.1 (3' sequence-TGTGCGAGAGA) with D1-1 (5' sequence-GGTACAACTGGAACGAC). N represents all the bases possible (ATGC). In other words, there is a need to design oligos for all the possibilities designated by N (4096 combinations). If only one base is inserted then the best signals will be from the top and bottom oligos (below). However, if two bases are inserted then the best signals will be from the first and second top and bottom oligos.

TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG
TGTGCGAGAGANNNNNNGGTACAACTGG

Example 7 - RNA and/or DNA extraction

DNA can be extracted from the sample according to means well known in the art (see, e.g., Maniatis, et al., Molecular Cloning; A Laboratory Manual (Cold Spring Harbor Lab, New York, 1982) or by using the RNA and/or DNA are extracted by commercial kits using the manufacturer's instructions (Amersham Pharmacia Biotech Ltd., Little Chalfont, UK) or by any other known in the art.

Example 8 - cDNA Synthesis and PCR Amplification.

It has shown by numerous studies that the region of the variable region of TCR and antibodies can be amplified using specific primers for RT-PCR and PCR. The amplification is by using sets of oligonucleotide primers that can primer all possible chain sequences. The oligonucleotide sequences of the 5' primers that are typically used are base on the N-terminal sequences of antibodies are the first framework determining region or the signal sequence. The sequences of the 3' primers are based on the

conserved regions of the first constant domain of antibodies or on the antibody J region sequences. A human Ig set of primers is available from Novagen. PCR is performed according to manufacture's protocol (Novagen, Germany).

Example 9 - Probe Labeling Procedures

a) PCR with end labeled fluorescent dye primers. The incorporation of the dye will be during the primer-annealing step. The 5' primers (for instance Cy3 labeled) will complement the leader, FR1, FR2, and FR3. This should be a mixture of primers that will anneal to all rearranged genes at a defined position within the gene. The 3' primer (for instance Cy5 labeled) is from the end of: FR2, FR3, FR4 (J region) or the beginning of the constant region.

b) PCR as described above, with non-labeled primer. Instead, the labeling is done by performing the PCR in the presence of labeled nucleotides.

c) In either case the above procedures can be done separately for each chain type and chain subtype. Or different type of labels can be used for each of the chains types.

End labeling of oligonucleotides is performed by using the 3' End Labeling Kit (NEN life Science Products, MA).

Example 10 - Probe Set (Chip Design)

The chip is composed of oligos in the range of 8-40 bases. Table 6 illustrates the number of overlapping oligos (with a single base shift) that are needed to cover the ~40,000 bases of the human germline gene fragments. The oligos numbers calculated in the table do not take into consideration the junction sequence that composes the recombinant variable region. It is evident that as the oligo length increases there is a decrease in the chance that a certain oligo sequence will occur more than once in a gene fragment.

Table 6. Statistics on the Oligo Frequency within the Ig Human Germline Gene Fragments

Oligo length	Number of oligos that are needed to span all germline segments*	Number of oligos that compose the unique set	Number of oligos that occur more than once in a gene segment	Number of oligos that occur in more than a single family (VH, VK, VL, D, J)
8	2766	2491	17	56
12	2734	2642	0	0
14	2718	2650	0	0

Example 11 - Determining Labeled Probe Level

After the chemical part of the experiment, it is subjected to image analysis. Data are collected by the scanner, digitized and stored. The task of the image processing hardware and software is to locate spots on the slide and to segment them, that is, to separate into the pixels bearing the signal and the pixels belonging to the background. There are several methods to perform this task as described in Holloway et al., Nat Genet suppl: 481-489, 2002 and Forster et al., J Endocrinol 178: 195-204, 2003.

Example 12 - Sequence Assembly Process.

A. gi|23320665|gb|AY056842.1| Homo sapiens anti-HIV-1 gp120 immunoglobulin heavy chain variable region mRNA, partial cds.

Frameworks regions are shown in yellow (FR1-FR4).
CDR regions are shown in red (CDR1-3).

```
CAGGTGCAGCTGGTGCAGTCTGGGGGAGGCCTAGTCCAGCCGGGGGGGTCCCTGAGACTCTCCTGTGCCGCCTTTGGATTCAACTTCAGT
TGGGTCCGCCAGGCTCCAGGGAAGGGACTGGAATATCTTTCA                                        AGATTCACCAGC
TCCAGAGACAATTCCAAGAACACACTGTTTCTTCAAATGGGCAGCCTGAGAACTGAGGACGTGGCTGTCTACTACTGTGCGAGA
          TGGGGCCAAGGAACAATGGTCGTCGTCTCTTC
```

B. Multiple alignment of AY056842.1 with: VH3-64, D3-22 and JH3.

```
                          .    :    .    :    .    :    .    :    .    :    .    :     60

VH3-64          GAGGTGCAGCTGGTGGAGTCTGGGGAAGGCTTGGTCCAGCCTGGGGGGTCCCTGAGACTC      60
GI|23320665|GB  CAGGTGCAGCTGGTGCAGTCTGGGGGAGGCCTAGTCCAGCCGGGGGGGTCCCTGAGACTC      60

                          .    :    .    :    .    :    .    :    .    :    .    :    120
VH3-64          TCCTGTGCAGCCTCTGGATTCACCTTCAGTAGCTATGCTATGCACTGGGTCCGCCAGGCT      120
GI|23320665|GB  TCCTGTGCCGCCTTTGGATTCAACTTCAGTTCCTATGTTATGCACTGGGTCCGCCAGGCT      120

                          .    :    .    :    .    :    .    :    .    :    .    :    180
VH3-64          CCAGGGAAGGGACTGGAATATGTTTCAGCTATTAGTAGTAATGGGGGTAGCACATATTAT      180
GI|23320665|GB  CCAGGGAAGGGACTGGAATATCTTTCAGCGATTAGTAGTGATGGAGAGACAACATATCAT      180

                          .    :    .    :    .    :    .    :    .    :    .    :    240
VH3-64          GCAGACTCTGTGAAGGGCAGATTCACCATCTCCAGAGACAATTCCAAGAACACGCTGTAT      240
GI|23320665|GB  GCAAACTCTGTGAAGGGCAGATTCACCAGCTCCAGAGACAATTCCAAGAACACACTGTTT      240

                          .    :    .    :    .    :    .    :    .    :    .    :    300
D3-22                                                                      GT       2
VH3-64          CTTCAAATGGGCAGCCTGAGAGCTGAGGACATGGCTGTGTATTACTGTGCGAGAGA          296
GI|23320665|GB  CTTCAAATGGGCAGCCTGAGAACTGAGGACGTGGCTGTCTACTACTGTGCGAGAGATCG-      299

                          .    :    .    :    .    :    .    :    .    :    .    :    360
D3-22           ATTACTATGATAGTAGTGGTTATTACTAC                                      31
JH3                                          GATGCTTTTGATGTCTGGGGCCAAGGGACA        30
GI|23320665|GB  -TTACTATGAGACTAGTGGTT-------CCAATGCTTTTGATGTCTGGGGCCAAGGAACA      351
```

```
                    .    :    .    :    .    :    .    :    .    :    .    :   420
JH3                 ATGGTCACCGTCTCTTCAG                                          49
GI|23320665|GB ATGGTCGTCGTCTCTTCA                                               369
```

C. Sequence from base 2-60 of AY056842.1 (framework 1). A match of AY056842.1 to VH3-64 is represented by a dot.

Sequence (probe) name	Position in gene	# Mutations (position in oligo,type)	Sequence
Germline- VH3-64	2-60		.....G.....A....T.G.....T.....
AY056842.1	2-60		AGGTGCAGCTGGTGCAGTCTGGGGGAGGCCTAGTCCAGCCGGGGGGGTCCCTGAGACTC
VH3-64	2-13	0	AGGTGCAGCTGG
VH3-64	3-14	0	GGTGCAGCTGGT
VH3-64	4-15	0	GTGCAGCTGGTG
VH3-64	5-16	1 (12,G ->C)	TGCAGCTGGTGC
VH3-64	6-17	1 (11,G ->C)	GCAGCTGGTGCA
VH3-64	7-18	1 (10,G ->C)	CAGCTGGTGCAG
VH3-64	8-19	1 (9 ,G ->C)	AGCTGGTGCAGT
VH3-64	9-20	1 (8 ,G ->C)	GCTGGTGCAGTC
VH3-64	10-21	1 (7 ,G ->C)	CTGGTGCAGTCT
VH3-64	11-22	1 (6 ,G ->C)	TGGTGCAGTCTG
VH3-64	12-23	1 (5 ,G ->C)	GGTGCAGTCTGG
VH3-64	13-24	1 (4 ,G ->C)	GTGCAGTCTGGG
VH3-64	14-25	1 (3 ,G ->C)	TGCAGTCTGGGG
VH3-64	15-26	2 (2 ,G ->C);(12,A ->G)	GCAGTCTGGGGG
VH3-64	16-27	2 (1 ,G ->C);(11,A ->G)	CAGTCTGGGGGA
VH3-64	17-28	1 (10,A ->G)	AGTCTGGGGGAG
VH3-64	18-29	1 (9 ,A ->G)	GTCTGGGGGAGG
VH3-64	19-30	1 (8 ,A ->G)	TCTGGGGGAGGC
VH3-64	20-31	2 (7 ,A ->G);(12,T ->C)	CTGGGGGAGGCC
VH3-64	21-32	2 (9 ,A ->G);(12,T ->C)	TGGGGGAGGCCT
VH3-64	22-33	3	GGGGGAGGCCTA *
VH3-64	23-34	3	GGGGAGGCCTAG *
VH3-64	24-35	3	GGGAGGCCTAGT *
VH3-64	25-36	3	GGAGGCCTAGTC *
VH3-64	26-37	3	GAGGCCTAGTCC *
VH3-64	27-38	2 (5 ,T ->C);(8 ,G ->A)	AGGCCTAGTCCA
VH3-64	28-39	2 (4 ,T ->C);(7 ,G ->A)	GGCCTAGTCCAG
VH3-64	29-40	2 (3 ,T ->C);(6 ,G ->A)	GCCTAGTCCAGC
VH3-64	30-41	2 (2 ,T ->C);(5 ,G ->A)	CCTAGTCCAGCC
VH3-64	31-42	3	CTAGTCCAGCCG *
VH3-64	32-43	2 (2 ,G ->A);(11,T ->G)	TAGTCCAGCCGG
VH3-64	33-44	2 (1 ,G ->A);(10,T ->G)	AGTCCAGCCGGG
VH3-64	34-45	1 (9 ,T ->G)	GTCCAGCCGGGG
VH3-64	35-46	1 (8 ,T ->G)	TCCAGCCGGGGG
VH3-64	36-47	1 (7 ,T ->G)	CCAGCCGGGGGG
VH3-64	37-48	1 (6 ,T ->G)	CAGCCGGGGGGG
VH3-64	38-49	1 (5 ,T ->G)	AGCCGGGGGGGT
VH3-64	39-50	1 (4 ,T ->G)	GCCGGGGGGGTC
VH3-64	40-51	1 (3 ,T ->G)	CCGGGGGGGTCC
VH3-64	41-52	1 (2 ,T ->G)	CGGGGGGGTCCC
VH3-64	42-53	1 (1 ,T ->G)	GGGGGGGTCCCT
VH3-64	43-54	0	GGGGGGTCCCTG
VH3-64	44-55	0	GGGGGTCCCTGA
VH3-64	45-56	0	GGGGTCCCTGAG
VH3-64	46-57	0	GGGTCCCTGAGA
VH3-64	47-58	0	GGTCCCTGAGAC
VH3-64	48-59	0	GTCCCTGAGACT
VH3-64	49-60	0	TCCCTGAGACTC

*** represents probes with three mutations. These probes might not be present on the chip if the maximum designed mutations was two. However, the sequence might still be assembled as shown in sections E.**

D. Probes that match the germline gene VH3-64. This is the sequence that is assembled using the full matching probes.

.....G.....A....T.G.....T.....
AGGTGCAGCTGGTGCAGTCTGGGGGAGGCCTAGTCCAGCCGGGGGGGTCCCTGAGACTC
AGGTGCAGCTGG
GGTGCAGCTGGT
GTGCAGCTGGTG
GGGGGGTCCCTG
GGGGGTCCCTGA
GGGGTCCCTGAG
GGGTCCCTGAGA
GGTCCCTGAGAC
GTCCCTGAGACT
TCCCTGAGACTC

E. Some probes in the overlapping scheme will not appear if the mutations designed will contain a maximum of two mutations. However, there is enough overlap to allow the sequence assembly.

Sequence (probe) name	Position in gene	# Mutations (position in oligo,type)	Sequence
Germline - VH3-64	1-59		.....G.....A....T.G.....T.....
AY056842.1	2-60		AGGTGCAGCTGGTGCAGTCTGGGGGAGGCCTAGTCCAGCCGGGGGGGTCCCTGAGACTC
VH3-64	20-31	2 (9 ,A ->G);(12,T ->C)	TGGGGGAGGCCT
VH3-64	26-37	2 (5 ,T ->C);(8 ,G ->A)	AGGCCTAGTCCA

F. Illustrated is the VDJ junction of sequence AY056842.1. All the probes that contribute to its assembly are shown.

Sequence (probe) name	Position in gene	# Mutations (position in oligo,type)	Sequence
AY056842.1	276		TGTCTACTACTGTGCGAGAGATCGTTACTATGAGACTAGTGGTT-------CCAATGCTTTTGATGTCTGGGGCCAAGGAACA
VH3-64	276		TGTGTATTACTGTGCGAGAGA
D3-22	1		GTATTACTATGATAGTAGTGGTTATTACTAC
JH3	1		GATGCTTTTGATGTCTGGGGCCAAGGGACA
VH3-64	276-287	2(4, G->C); (7, T ->C)	TGTCTACTACTG
VH3-64	277-288	2(3, G->C);	GTCTACTACTGT

		(6, T ->C)	
VH3-64	278-289	2(2, G->C); (5, T ->C)	TGTACTACTGTG
VH3-64	279-290	2(1, G->C); (4, T ->C)	CTACTACTGTGC
VH3-64	280-291	1(3, T ->C)	TACTACTGTGCG
VH3-64	281-292	1(2, T ->C)	ACTACTGTGCGA
VH3-64	282-293	1(1, T ->C)	CTACTGTGCGAG
VH3-64	283-294	0	TACTGTGCGAGA
VH3-64	284-295	0	ACTGTGCGAGAGA
8MER			CGAGAGAT
8MER			GAGAGATC
8MER			AGAGATCG
8MER			GAGATCGT
8MER			AGATCGTT
8MER			GATCGTTA
8MER			ATCGTTAC
8MER			TCGTTACT
8MER			CGTTACTA
D3-22	1-12	2(1 ,A ->G); (11,T ->G)	GTTACTATGAGA
D3-22	2-13	2(10,A ->G); (12, G ->C)	TTACTATGAGAC
D3-22	3-14	2(9, A ->G); (11, G ->C)	TACTATGAGACT
D3-22	4-15	2(8, A ->G); (10, G ->C)	ACTATGAGACTA
D3-22	5-16	2(7, A ->G); (9, G ->C)	CTATGAGACTAG
D3-22	6-17	2(6, A ->G); (8, G ->C)	TATGAGACTAGT
D3-22	7-18	2(5, A ->G); (7, G ->C)	ATGAGACTAGTG
D3-22	8-19	2(4, A ->G); (6, G ->C)	TGAGACTAGTGG
D3-22	9-20	2(3, A ->G); (5, G ->C)	GAGACTAGTGGT
8MER			TAGTGGTC
8MER			AGTGGTCC
8MER			GTGGTCCA
8MER			TGGTCCAA
8MER			GGTCCAAT
8MER			GTCCAATG
8MER			TCCAATGC
JH3	1-12	2(1, G ->C); (3, G ->A)	CCAATGCTTTTG
JH3	2-13	1 (2, G ->A)	CAATGCTTTTGA
JH3	3-14	1 (1, G ->A)	AATGCTTTTGAT
JH3	4-15	0	ATGCTTTTGATG
JH3	5-16	0	TGCTTTTGATGT
JH3	6-17	0	GCTTTTGATGTC
JH3	7-18	0	CTTTTGATGTCT
JH3	8-19	0	TTTTGATGTCTG
JH3	9-20	0	TTTGATGTCTGG
JH3	10-21	0	TTGATGTCTGGG
JH3	11-22	0	TGATGTCTGGGG
JH3	12-23	0	GATGTCTGGGGC
JH3	13-24	0	ATGTCTGGGGCC
JH3	14-25	0	TGTCTGGGGCCA
JH3	15-26	0	GTCTGGGGCCAA
JH3	16-27	0	TCTGGGGCCAAG
JH3	17-28	0	CTGGGGCCAAGG
JH3	18-29	1 (12, G ->A)	TGGGGCCAAGGA
JH3	19-30	1 (11, G ->A)	GGGGCCAAGGAA
JH3	20-31	1 (10, G ->A)	GGGCCAAGGAAC
JH3	21-32	1 (9, G ->A)	GGCCAAGGAACA

Example 13 - All Possible Oligonucleotides Derived from VH3-64 (Base 1 to 18) that encode for the first six amino acids.

Sequence (probe)/ name	Position in gene	Amino acid sequence	Base Sequence
Germline VH3-64	1-18	E V Q L V E	GAGGTGCAGCTGGTGGAG
VH3-64	1-12	E V Q L	GA[GA]GT[ATGC]CA[AG][CT]T[ATGC]
VH3-64	2-13	V Q L	A[GA]GT[ATGC]CA[AG][CT]T[ATGC]G
VH3-64	3-14	V Q L	[GA]GT[ATGC]CA[AG][CT]T[ATGC]GT
VH3-64	4-15	V Q L V	GT[ATGC]CA[AG][CT]T[ATGC]GT[ATGC]
VH3-64	5-16	Q L V	T[ATGC]CA[AG][CT]T[ATGC]GT[ATGC]G
VH3-64	6-17	Q L V	[ATGC]CA[AG][CT]T[ATGC]GT[ATGC]GA
VH3-64	7-18	Q L V E	CA[AG][CT]T[ATGC]GT[ATGC]GA[GA]

While the present invention has been particularly described, persons skilled in the art will appreciate that many variations and modifications can be made. Therefore, the invention is not to be construed as restricted to the particularly described embodiments, rather the scope, spirit and concept of the invention will be more readily understood by reference to the claims which follow.

CLAIMS

1. A method of sequencing a plurality of nucleic acids in a mixture of unknown sequences comprising the steps of:

   (a) using conditions which differentiate an exactly complementary oligonucleotide probe and an oligonucleotide probe having at least one mismatched nucleotide;

   (b) contacting a plurality of oligonucleotides, each from eight to forty nucleotides in length, with said target nucleic acids;

   (c) forming duplexes between the plurality of target nucleic acids and the plurality of oligonucleotides;

   (d) washing the duplexes;

   (e) detecting oligonucleotides positively hybridizing as part of said duplexes; and

   (f) compiling a set of sequences of the plurality of nucleic acids from overlapping positively hybridizing oligonucleotides.

2. The method of claim 1 further comprising means for distinguishing the results obtained for different sequences in the target nucleic acids.

3. The method of claim 2 wherein the means for distinguishing the results comprise an algorithm for analyzing the data obtained form the hybridizations.

4. The method of claim 3 further comprising means for storing the data in a data storage system.

5. The method of claim 1 wherein the mixture of target nucleic acids is selected from transcripts of immunoglobulin genes and T cell receptor genes.

6. The method of claim 5 wherein the mixture of target nucleic acids are obtained from a single individual.

7. The method of claim 6 wherein the stored data further comprise information correlated to the health or disease state of the individual from whom the target nucleic acids were obtained.

8. A method for partial sequencing of a plurality of nucleic acids, comprising the steps of:

(a) using conditions which differentiate an exactly complementary oligonucleotide probe and an oligonucleotide probe having at least one mismatched nucleotide;

(b) contacting the plurality of target nucleic acids in a reaction mixture with a plurality, but less than a totality, of oligonucleotide probes of given length, each at least eight nucleotides in length;

(c) forming duplexes between the target nucleic acids and the plurality of oligonucleotides;

(d) washing the duplexes;

(e) detecting oligonucleotides positively hybridizing as part of said duplexes; and

(f) compiling a partial sequence of said target nucleic acids from a subset of said oligonucleotide probes which form duplexes with said target nucleic acid and which overlap with at least one other member of said oligonucleotide probes.

9. The method according to claim 8, wherein said oligonucleotide probes are each from eight to forty in length.

10. A method of identifying nucleic acid probes to quantify the expression of a plurality of target nucleic acids, comprising:

(a) providing a high density array of nucleic acid probes, said array comprising a multiplicity of nucleic acid probes, wherein each probe is complementary to a subsequence of a target nucleic acid and for each probe there is at least one corresponding mismatch control probe;

(b) hybridizing said target nucleic acid to said array of nucleic acid probes; and

(c) identifying those probes for which the difference in hybridization signal intensity between each probe and its mismatch controls is detectable.

11. A method of analyzing the expression of one or more genes encoding antibodies or antibody fragments and T cell receptors or T cell receptor fragments, comprising:

(a) providing a pool of target nucleic acids comprising RNA transcripts of one or more of said genes, or nucleic acids derived therefrom using said RNA transcripts as templates;

(b) hybridizing said pool of nucleic acids to an array of nucleic acid probes immobilized on a surface, said array comprising at least 100 different nucleic acids, at least some of which comprise control probes, wherein each different nucleic acid is localized in a known location on said surface, and at least some of said nucleic acid probes are complementary to said RNA transcripts or said nucleic acids derived therefrom using said RNA transcripts; and

(c) quantifying the hybridization of said nucleic acids to said array by comparing binding of matched and control probes.

12. The method of claim 11, wherein said control probes are mismatch control probes.

13. The method of claim 12, wherein said quantifying comprises calculating the difference in hybridization signal intensity between each of said nucleic acid probes and its corresponding mismatch control probes.

14. The method of claim 12, wherein said quantifying comprises calculating the difference in hybridization signal intensity between each of said nucleic acid probes and the intensity of a background signal.

15. The method of claim 11, wherein for each gene, said array comprises at least 10 different nucleic acid probes complementary to subsequences of the gene.

16. The method of claim 11, wherein said target nucleic acids are extracted from cells of the immune system existing in any suitable specimen or sample, selected from a group consisting of blood, tissue samples, selected cell subpopulations bone marrow, lymph nodes, thoracic duct, Peyer's patch or any other organ or bodily fluid.

ABSTRACT

The present invention relates to the sequencing of expressed genes belonging to the immunoglobulin gene superfamily, particularly immunoglobulins and T cell receptors. In particular, the present invention provides methods of sequencing expressed genes in a non-clonal population of cells of the immune system, generating profiles of the genes expressed, and correlating the data generated with states of disease or health.

Figure 1/5

Flowchart of the invention process

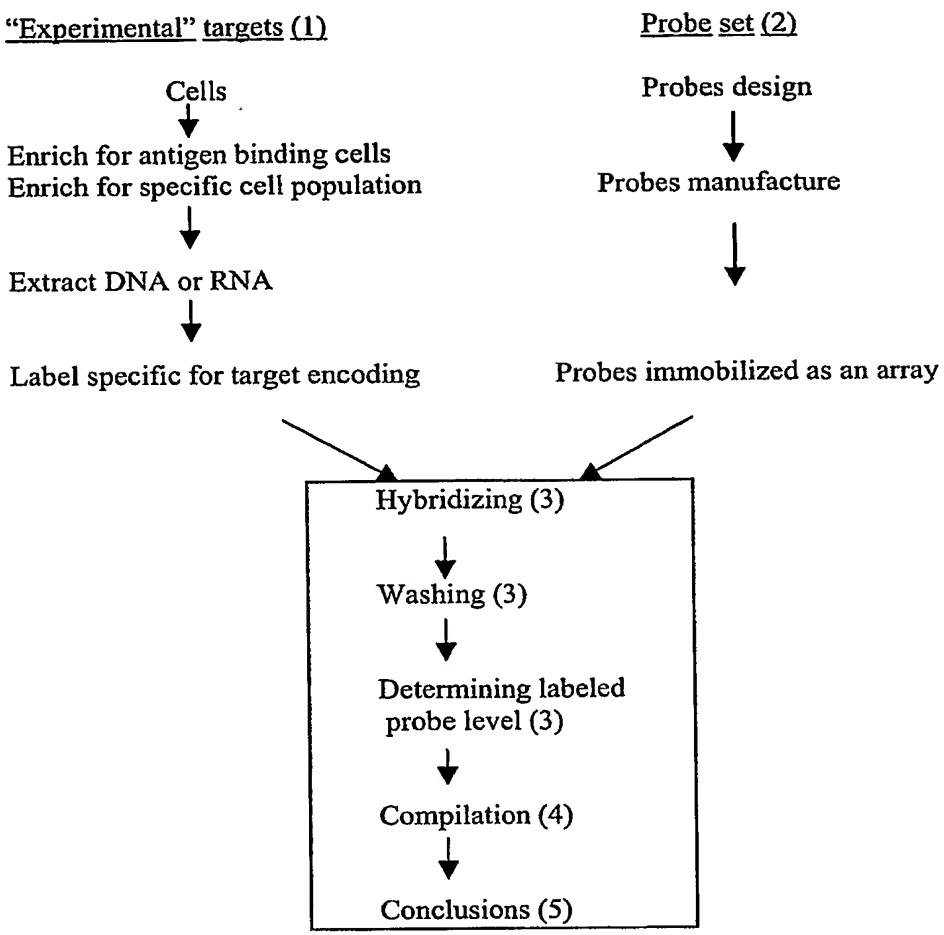


Figure 2/5

Illustration of the variable region gene and an example of overlapping positively labeled probes

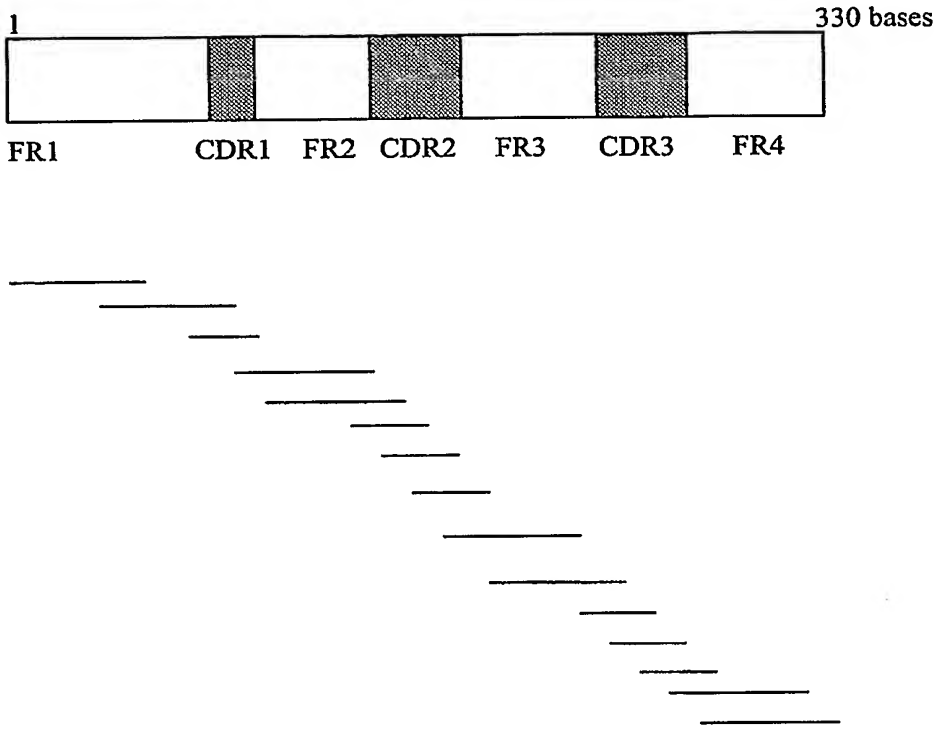
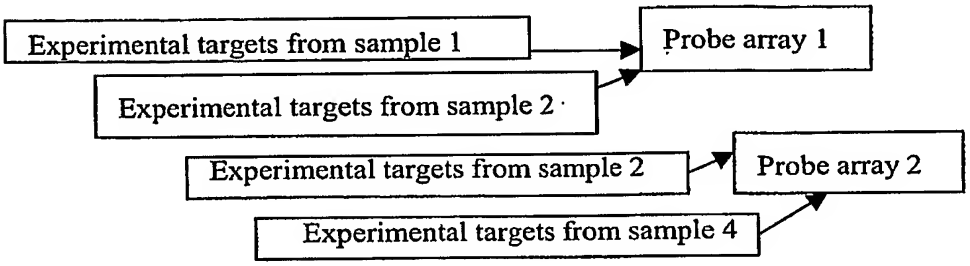


Figure 3/5

Examples of multiple sample experiment. The probe arrays represented have the same content.

A.



B.

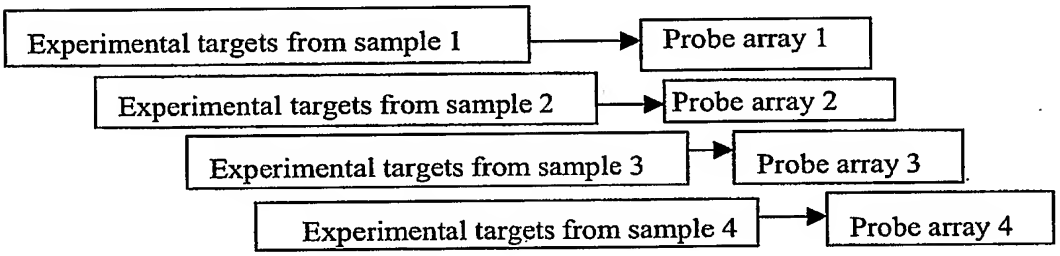


Figure 4/5

Scheme illustrating end labeled plus and minus strand.

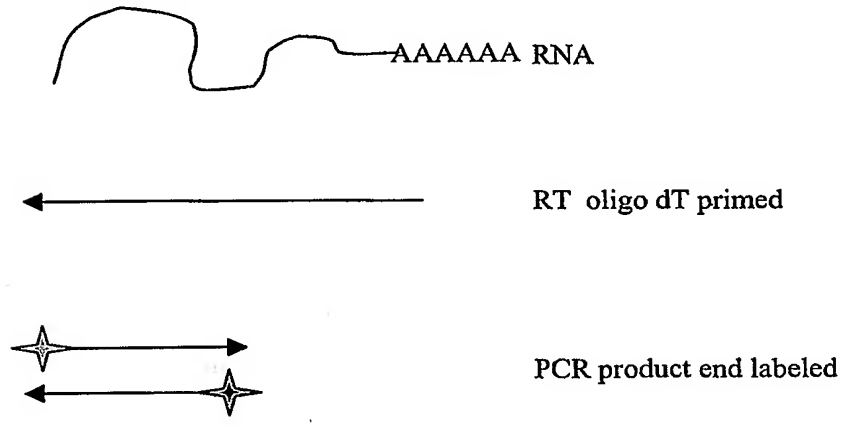


Figure 5/5

Scheme illustrating PCR amplification of specific regions within the target sequence.